***Istat’s new strategy and tools for enhancing statistical utilization of the available administrative databases***

*Giovanna D’Angiolini, Pierina De Salvo, Andrea Passacantilli – Italian National Institute of Statistics (ISTAT)*

***Abstract*** The paper presents the Istat’s strategy for supporting the statistical users in exploiting administrative data sources. In particular Istat is launching several activities aimed at producing standard documentation about the administrative data sources’ information content and quality, which is managed and disseminated to users through a dedicated web-based metadata management system, called DARCAP. Moreover in order to support in-depth quality analyses we are studying a new Framework of Quality Indicators for Administrative Data Sources.

***Introduction***

Istat is undertaking a general strategy aimed at making the administrative data sources the more usable as possible for statistical purposes by means of defining proper activities and dedicated tools and levering the collaboration with the owner institutions [1].

In order to exploit administrative data sources for statistical purposes, the first step is the characterization of their content in terms of collectives which may be of interest to the statisticians, with their features. For this purpose it is important to regard the administrative procedures as data collecting tools and characterize their collected data as pieces of information about real-world observable items which may assume a potential interest for statistical users. The aim is to single out what can be statistically used in a given administrative data source, independently of any current utilization, from a strictly documentation viewpoint. This requires an appropriate description of the administrative data source’s information content, namely the definition of the administrative data source’s ontology. This goal implies the capability of specifying the content of the administrative data source in terms of several collectives with their features, taking into account that an administrative data source generally observes both populations and sets of events which occur in time.

The more the statistical users exploit existing data sources, in particular administrative data sources, the more the description of the ontology of these sources in a standard and understandable way becomes important, independently of any further particular statistical use. Moreover, the statistical users need an accurate and comparable assessment of the data source’s quality, which should exactly concern those collectives and features they are interested in.

The Istat’s strategy is carried out through activities aimed at the specifying the information content of each administrative data source and analysing and measuring its quality, and through the Istat’s supervision on those changes and innovation projects which involve administrative data sources and administrative forms. In particular the specification of each administrative data source’s content and quality is attained by means of investigations on administrative data sources and their related administrative forms. An investigation is an analysis and documentation activity which employs standard tools and is undertaken by Istat in collaboration with the owner institution. In order to support such activities Istat is building methodological and information managing tools, namely the DARCAP system and the Quality Assessment Framework for Administrative Data Sources.

*DARCAP (Documenting ARChives of Public Administrations)* is the web-based information management system for supporting the administrative data sources’ investigations and other documentation initiatives in order to provide the administrative data sources’ potential users with structured documentation of their content and features. This tools also supports the administration institutions in sending Istat their communications about the innovation initiatives which concern administrative data sources or administrative forms. Furthermore DARCAP supports Istat experts in producing structured documentation of the new information content of the administrative data sources which are involved in innovation projects, and in defining Istat’s recommendations.

The *Quality Assessment Framework for Administrative Data Sources* is the Istat’s methodological tool for supporting the statistical users in evaluating the quality of the available administrative data sources.

***Documenting the content and the quality of the available administrative data sources***

The investigation on administrative data sources is performed by means of analyzing the available documentation and interviewing the source’s experts belonging to the owner institution as well as the source’s users. The collected documentation is then structured according to the DARCAP’s database structure, in order to be stored into such a database.

First we specify the denomination and the main characteristics of the administrative data source, the owner and the other managing institutions, those information flows and sets of administrative forms which are currently used for feeding the administrative data source with data. Furthermore the administrative data source’s content is documented, in particular the main observed populations, which correspond to those collectives which are the target of the administrative procedure, and their related sets of events, each one with its definition. We also document the main characteristics which are possessed by the single elements belonging to the specified collectives with their definitions, and the associated classifications (list of modalities) for qualitative characteristics. This work of conceptual description of the administrative data source’s content produces a source ontology’s specification which encompasses: the main collectives, which may be populations or sets of events, the main characteristics of these populations or sets of events, and also the relationships which link populations and sets of events. The first result of this work is a network of main populations or sets of events, linked by 1-1 or 1-N relationships, in which every collective has its own definition and characteristics. A further analysis of the administrative data source leads to single out more populations or sets of events which have associated their distinguished characteristics and relationships and are linked with the main collectives by means of subset relationships.

By means of the investigation on administrative data sources we also produce a first evaluation of their quality. More precisely we ask the source’s experts for information concerning each population or set of events. For each population we document the enter and exit events and the way by which their registration influence the population’s coverage. For each set of events we document the way by which the single events are recorded into the source and the time distribution of events as well as the coverage problems due to: registration scope, namely the capability of effectively registering all the single expected events, registration systematic distortion related to the purposes of the administrative registration procedure, registration timeliness, namely the time lag between the occurrence of the event and its registration. The main problems and the possible interventions concerning the collectives’ definitions, the suitability of the used classifications and their correspondence with standard classifications, the identification codes which may be used for the exact linking with other data sources are also evaluated. For the administrative data source as a whole, the main problems and the possible interventions concerning its statistical usability and its diffusion timeliness are also evaluated together with the related innovation strategies.

For the purpose of a deeper analysis of the quality of administrative data sources it is useful and necessary to calculate standard quantitative indicators. As described in the next dedicated section the Framework of quality indicators for administrative data sources defines concepts, methods and specific indicators for such an in-depth quality evaluation.

The investigation activity may suggest how to improve the content and the quality of the investigated sources. Moreover in order to enable Istat to realize a more direct intervention on the existing administrative data sources we are now launching another activity, namely the supervision on changes and innovation projects related to the observed administrative data sources. In order to accomplish this task, Istat is asking a first group of administrative data sources’ owner institutions to inform Istat about any kind of innovation project concerning their administrative forms and data sources in order to receive a technical and scientific evaluation. The DARCAP system provides a dedicated environment for collecting such communications, which may concern occasional as well as periodic changes, analyzing them to a certain extent, storing documentation about the communicated innovation projects as well as about the whole process of analyzing them and releasing opinions and recommendations.

All the above described activities are coordinated by a *Committee for Harmonizing Administrative Forms* whose members are nominated by Istat and the most important administrative data sources’ owner institutions, which is supported by a *Network of experts*.

***Documenting the content of administrative data sources: the conceptual model***

The documentation activity aims at producing a standard and therefore comparable specification of the content of the available administrative data sources in terms of observed real-world objects, namely an ontology of the documented administrative data sources.

An ontology of an administrative data source is a structured description of its information content, based on a standard conceptual model. In order to define such a conceptual model, we have analyzed the life-cycle of the administrative data and singled out the different kinds of real-world objects to which they are referred, and we have put such objects into correspondence with those objects to which any statistic is currently referred, namely collectives and variables. Our conceptual model is oriented towards supporting the statistical exploitation of the administrative data sources, but it can be easily translated into other general-purpose conceptual models and languages for ontology specification [2]. In the following we briefly introduce its main features.

Administrative data sources collect information about several kinds of real world objects in order to support administrative activities [3]. First, any administrative activity entails collecting data about those entities which the activity addresses. Such entities are subsets of the two general populations of persons, on one side, and entities which perform economic activities, on the other side, or they are subsets of related populations such as households, territorial units. Moreover, information is collected about those particular sets of events which may involve these entities and are of interest for the purposes of the administrative activity. The observed *populations* and *sets of events* are linked by r*elationships*. For both observed populations and sets of events proper information is collected about their characteristics, which may change in time. As an example, the Ministry for Public Education continuously collects information about the students, the schools and the universities with their characteristics as well as about sets of events such as the degree course enrolments, the examinations, the degree earnings with their characteristics.

Therefore inside an administrative data source we find two kind of linked collectives: *populations* and *set of events*. Populations are subsets of the two most general populations of persons on one side, and entities which perform economic activities on the other side, or subsets of their related populations. Sets of events can be instantaneous (such as examination) or durable (such as degree course enrolment) and they may connect elements belonging to different populations, as an example any degree course enrolment event connects a student with a degree course. Each element of these collectives has *qualitative* or *quantitative* *characteristics*, such as date of birth, residence, date of the enrolment, examination score, as well as relationships with elements in other collectives.

According to a widespread ontology specification paradigm, in our conceptual model a qualitative or quantitative characteristic is regarded as a relation which links an element belonging to a collective with an item belonging to a proper *classification*, or with a number in a numerical domain respectively. From a statistical viewpoint, the quantitative characteristics and the qualitative characteristics together with their associated classifications are regarded as variables. New variables can be defined as combinations of relationships and characteristics by means of logical and numerical operators, this is the reason why it is important to document the relationships among collectives. Finally an administrative data source’s ontology is a network of populations and sets of events which are linked by 1-1 or 1-n relationships and have associated quantitative or qualitative characteristics, the latter ones with their associated classifications.

Often some characteristics or relationships are associated with only a part of the elements of a collective. In this case it is worth to define another collective which is a subset of the main collective, whose elements have associated such characteristics or relationships. More precisely, we distinguish between *subset relationships* and *partition relationships*. A subset relationship simply links two collectives when one gathers a part of the elements of the other. A partition relationship links a collective with many collectives which jointly partition it, that is: each element of the partitioned collective belongs to one and only one of the partitioning collectives.

***Assessing the quality of administrative data sources: building the Framework of quality indicators for administrative data sources***

Trends such as the open data vision, the widespread development of data warehouses, the increasing usage of administrative data sources for statistical purposes not only by NSOs, but also by other organizations including their owner organizations’ themselves, are all factors which are enlarging the scope of the quality assessment activity. In this scenario the NSOs should take responsibility for a new methodological coordination task, namely to define a rich and flexible enough sets of standard and repeatable quality assessment procedures for administrative data sources, as they currently do for surveys [1]. In order to meet such a requirement we have based our Framework on a careful analysis of the particular goals and features of the administrative data collection process and their effects on the quality of the collected data for each one of the different kinds of observed objects which set up any data source’s ontology [4].

Our Framework is organized according to the structure that has been proposed by Statistics Netherlands [5], which distinguishes three different views on quality, namely the Source view, the Metadata view, and the Data view. To each of these views called “hyperdimension” is associated a number of dimensions, quality indicators and methods.

In the *Source hyperdimension*, the quality aspects relate to the administrative data source as a whole, the data set keeper, and the delivery conditions. The *Metadata hyperdimension* specifically focuses on the metadata related aspects of the administrative data source. It is concerned with the existence and the adequacy of the documentation and with the kind and the structure of the identification codes. The *Data hyperdimension* focuses on the quality aspects of the data in the administrative data source. For the Source and Metadata hyperdimensions we propose a set of qualitative indicators. As we have seen, in addition to requiring the administrative data owners to certify the availability of proper metadata we provide them with a standard tool for metadata specification, namely the DARCAP system.

As to the indicators in the Data hyperdimension, according to our approach we aim to define a rich and well-reasoned quality indicators’ frame in order to drive anyone outside or inside a NSO, particularly the administrative data source’s owners themselves, in calculating and interpreting each indicator, therefore *the quality indicators are defined on the basis of the data set’s ontology specification* and *the Data hyperdimension includes both qualitative and quantitative indicators*.

As we have seen the qualitative indicators in the Data hyperdimension are specified by asking the data set experts a first qualitative assessment concerning some preliminary aspects of the data quality, such as coverage and influence of registration delay on coverage, distinctly for each collective (populations and set of events) in the administrative data source. As to the quantitative indicators, namely those indicators which are calculated from data and therefore require the availability of the data set, they must be calculable by the administrative data owner as well as by the NSO when it acquires the data set, the best scenario is when a collaborative calculation procedure is applied.

*In order to define such quantitative indicators, first we have discriminated between possible errors, on one side, and ways of checking them, on the other side*. *The possible errors are defined in terms of those objects that may be present in an administrative data source’s ontology*, in the following way.

For each object in an ontology, namely a collective, a characteristic, or a relationship we can build belonging statements concerning observed elements, more precisely we can assert that a single observed element belongs to a set or that a couple of elements belongs to a characteristic or a relationship. In logical terms, statements concerning populations and sets of events correspond to a single variable predicate, statements concerning characteristics and relationships correspond to two variable predicates. Such statements will be true or false.

As an example, let us suppose we have an administrative data source whose ontology encompasses:

* Student (x), Degree\_course (x), Examination (x) and Enrolment (x) which are collectives, more precisely Student (x) and Degree\_course (x) are populations, Examination (x) and Enrolment (x) are set of events
* Residence (x y), a characteristic which link each element x of the population Student (x) with an item y in the classification Town\_codes (y)
* Examination\_Student (x y) a relationship which links each element x of the set of events Examination (x) with an element y of the population Student (x)
* Enrolment\_Student (x y) and Enrolment\_Course (x y), two relationships which link each element of the set of events Enrolment (x) with an element y of the population Student (x), or Degree\_course (x) respectively.

Examples of belonging statements which involve observed elements are:

* the person identified by the fiscal code n is a student, namely Student (n)
* this person lives in Milan, namely Residence (n, Milan)
* there is an event i belonging to the set of the Examination events which concern such a person, namely Examination (i), Examination\_Student (i, n),
* there is an event i belonging to the set of the Enrolment events which concern such a person and the degree-course Statistics, namely Enrolment (i), Enrolment\_Student (i, n), Enrolment\_Course (i, Statistics).

However our conceptualization is more complex because the above statements have also time references as parameters, which are single moments for instantaneous events such as Examination, or (possibly open) periods for elements of populations such as Student, or durable events such as Enrolment.

The administrative data sources continuously collect and store data which are in fact proper combinations of such belonging statements. Referring to the above example, for each new student a new record is stored in the Student register which combines the statement Student (n) with the statement Residence (n, Milan) and others similar statements, and another new record is stored in the Student register too, which combines the statement Enrolment (i), with the statements Enrolment\_Student (i, n), Enrolment\_Course (i, Statistics) and possibly other statements.

Each administrative data source has its own data collection procedure, which consists in accepting or not accepting such combinations of belonging statements inside the data source. As a result, at any time any administrative data source stores a collection of belonging statements for each collective, characteristic or relationship in its ontology. It may happen that some of these statements are false, and that some true statements are not in the data set.

Therefore at any given time we may have in the administrative data source:

* *Inclusion errors*: false statements (definitely or temporarily) accepted in the data source
* *Exclusion errors*: true statements (definitely or temporarily) excluded from the data source

Other errors may concern *wrong identification of the involved elements*, because of problems in the identification code system, such as: syntactical errors in identifiers, identifiers for not existing elements, lack of identifiers for existing elements, more than one identifier for each element, elements which share identifiers.

For *each collective* (population or set of events) the inclusion or exclusion errors correspond to the well-known *over-coverage* and *under-coverage* errors respectively, and by means of combining them with identification errors we obtain a specification of all possible errors which concern belonging to the collective.

For *each mandatory characteristic* we may have an exclusion error, which corresponds to a nonresponse error, as well as a combined exclusion and inclusion error if the element is linked with a wrong item in the classification or a wrong numerical value, which corresponds to a *measure error*; for not mandatory characteristics we may have inclusion errors too. Identification errors may affect the observed characteristics too, when a change in a characteristic is registered for an element which is already in the data set, such as a new residence town for a student.

Errors which may concern relationships are specified in a similar way.

The available *quality check methods* are mainly: searching evident errors, such as duplicate identification codes, linking with other data sources, using logical constraints (mandatory or incompatible combinations between various belonging statements), calculating time lags between the moment of the events’ occurrence and the moment of their registration.

Until now we have defined a quality indicators’ frame concerning the collectives’ coverage and the elements’ identification by means of properly combining possible errors and quality checks methods. We are now analyzing the possible errors on characteristics and relationships in order to define two other quality indicators’ frames concerning all kinds of nonresponses, measure errors, relationship errors. Obviously in such indicators’ frames each quantitative indicator refers to several kinds of possible errors, therefore the quality evaluator will exploit the collected qualitative information about the nature and the respective relevance of the different kinds of errors in order to choose those quantitative indicators which are worth calculating and properly interpret them. Note that *our proposed indicators are distinctly calculable for each collective, characteristic and relationship in the administrative data source’s ontology, in order to effectively support any statistical usage of the collected information*.

***Conclusions***

After a one year long testing activity, *we are now ready to start the investigation activity and the supervision activity on innovation projects concerning administrative data sources at operating speed*. Thanks to the work done in this first experimentation year it has been possible to improve the investigations supporting tools and the applied work procedures in collaboration with the administrative data sources’ owner institutions.

We are now carrying on our work of specifying indicators in the Data hyperdimension on the basis of a careful analysis of possible errors anchored to the objects which may be present in an administrative data source’s ontology. We plan to integrate the existing indicators, such as the BLUE-ETS indicators [6], in our indicators’ framework as summaries of these more detailed indicators. Finally the *Framework of Quality indicators for Administrative Data Sources* will contain *qualitative indicators for a preliminary quality assessment* in the Source and Metadata hyperdimensions together with *a rich set of both qualitative and quantitative indicators for an in-depth and customizable quality assessment* in the Data hyperdimension, which can be summarized in a more limited frame for a first overall certification task.

Moreover in our opinion an important advantage of our approach is that it gives foundations for *future research aimed at building a generalized probabilistic frame for the quality assessment activity*, by means of properly reformulating the task of assessing the quality of any data collection as a problem of evaluating and composing the probabilities of all the possible errors.

***References***

[1] United Nations Economic Commission for Europe (UNECE). (2011), Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices, United Nations Publication

[2] D’Angiolini G., (2013), Manuale per la documentazione di archivi, moduli e dataset nel sistema DARCAP, Istat document

[3] Brackstone G.J. (1987), Issues in the use of administrative records for statistical purposes, Survey methodology

[4] D’Angiolini G., De Salvo P., Passacantilli A., Pogelli F., (2013), Framework per la qualità degli archivi amministrativi, Istat document

 [5] Daas, P., Ossen S., Vis-Visschers, R., and Arends-Tóth, J. (2009), Checklist for the Quality evaluation of Administrative Data Sources, Discussion paper by Statistics Netherlands

[6] Daas, P., Ossen S., Tennekes, M., Zhang , L., Hendriks , C., Foldal Haugen K., Cerroni, F., Di Bella G., Laitila, T., Wallgren, A., A.Wallgren, A., (2011) BLUE – ETS Deliverable 4.2 - Report on methods preferred for the quality indicators of administrative data sources***.***